What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech

Darina Benikova(✉), Michael Wojatzki, and Torsten Zesch

Language Technology Lab, University of Duisburg-Essen, Duisburg, Germany
{darina.benikova,michael.wojatzki,torsten.zesch}@uni-due.de

Abstract. We analyze whether implicitness affects human perception of hate speech. To do so, we use Tweets from an existing hate speech corpus and paraphrase them with rules to make the hate speech they contain more explicit. Comparing the judgment on the original and the paraphrased Tweets, our study indicates that implicitness is a factor in human and automatic hate speech detection. Hence, our study suggests that current automatic hate speech detection needs features that are more sensitive to implicitness.

1 Introduction

With the rise of social media, hate speech (HS) has moved into the focus of public attention. However, as its perception depends on linguistic, contextual, and social factors (Stefanowitsch 2014), there is no consensus on what constitutes HS. We examine a specific dimension of this challenge – whether implicitness affects HS perception. Consider the following Tweets:

Im. Everything was quite ominous with the train accident. Would like to know whether the train drivers were called Hassan, Ali or Mohammed #Refugee Crisis
Ex. Everything [...] - The train drivers were Muslims. #RefugeeCrisis

One could argue that the first Tweet is more offensive, since it evokes racist stereotypes by using allegedly prototypical Muslim first names as an implicit way of blaming Muslims in general. However, one could counter-argue that the second Tweet is more offensive, as it explicitly accuses Muslims of being involved in a train accident. Additionally, the first Tweet is hedged by *Would like to know whether*, whereas it is implied that the second statement is rather factual. It remains unresolved whether implicit or explicit HS is perceived as more offensive and what the role of hedging is (Sanchez and Vogel 2013).

In addition to the influence on the perception of HS, implicitness is a challenge for automatic HS detection, as most approaches rely on lists of abusive terms or phrases (Waseem and Hovy 2016).

Or in terms of the above example, the classifier learns that it is HS to agitate against *Muslims*, but fails to learn the connection to *Hassan*.

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 171–179, 2018.
https://doi.org/10.1007/978-3-319-73706-5_14

To shed light on the influence of implicitness on the perception of HS, we construct a dataset[1] in which we can experimentally control for implicitness. We select implicit HS instances from the German Hate Speech Twitter Corpus (Ross et al. 2016) and create explicit paraphrased counterparts[2]. We then conduct a user study, wherein we ask participants to rate the offensiveness of either implicit or explicit Tweets. We also show that a supervised classifier is unable to detect HS on both datasets.

We hypothesize that there is a measurable difference in the perception of implicit and explicit statements in both human and automatic performance. However, we cannot estimate the direction of the difference.

2 Theoretical Grounding

Our work is grounded in (i) research on detecting HS, (ii) the annotation and detection of implicit opinions, and (iii) on paraphrasing.

Detecting Hate Speech. Hitherto, there has been no work on HS detection considering the issues posed by implicitness. Approaches based on n-grams or word lists, e. g., (Sood et al. 2012; Chen et al. 2012) are limited to detecting explicit insults or abusive language. Methods involving more semantics, e. g., by incorporating Brown clusters (Waseem and Hovy 2016; Warner and Hirschberg 2012) are unlikely to cope with implicitness, as the necessary inferences go beyond word-relatedness.

Implicit Opinions. If we define HS as expressing a (very) negative opinion against a target, there is a clear connection to aspect-based sentiment analysis. However, sentiment analysis usually only models explicit expressions. For instance, the popular series of SemEval tasks on detecting aspect based sentiment, intentionally exclude implicit sentiment expressions and expressions requiring co-reference resolution in their annotation guidelines (Pontiki et al. 2014, 2015, 2016). Contrarily, the definition of stance, namely being in favor or against a target (i. e., a person, a group or any other controversial issue) explicitly allows to incorporate such inferences (for annotation guidelines see Mohammad et al. (2016) or Xu et al. (2016)). Thus, HS can also be considered as expressing a hateful stance towards a target.

Consequently, we define explicit HS as expressing hateful sentiment and implicit HS as the instances which do not express hateful sentiment, but hateful stance. Therefore, this work relates to studies which use explicit opinion expressions to predict or rationalize stance (Boltužić and Šnajder 2014; Hasan and Ng 2014; Sobhani et al. 2015; Wojatzki and Zesch 2016).

[1] https://github.com/MeDarina/HateSpeechImplicit.

[2] All examples in this paper are extracted from this corpus and were translated to English. None of the examples reflects the opinion or political orientation of the authors.

Paraphrasing. The implicit and explicit versions of a Tweet can be seen as paraphrases, i. e., units of texts containing semantically equivalent content (Madnani and Dorr 2010). Paraphrases can be classified according to the source of difference between the two texts. Incorporating implicit stances is equivalent to the paraphrase class of *Ellipsis* or the *Addition/Deletion* class.

The modification of hedges corresponds to the classes of *Quantifiers* and *General/Specific substitution* (Bhagat and Hovy 2013; Rus et al. 2014; Vila et al. 2014). To the best of our knowledge, paraphrasing techniques have not been used in the context of HS and its analysis.

3 Manufacturing Controllable Explicitness

The basis of our data set is the German Hate Speech corpus (Ross et al. 2016) that contains about 500 German Tweets annotated for expressing HS against refugees or not. We chose this corpus because it is freely available and addresses a current social problem, namely the debate on the so-called *European refugee crisis*. To construct a data set in which we can control for implicitness, we perform the following steps: (1) Restriction to Tweets which contain HS, i. e., at least one annotator flagged a Tweet as such (2) Removal of Tweets containing explicit HS markers, as described in Sect. 3.1 (3) Paraphrasing the remaining Tweets to be explicit, so that we obtain a dataset which has both an implicit and an explicit version of each Tweet.

3.1 Indicators for Explicit Hate Speech

We first identify tokens that are clear indicators for HS by retrieving words that are most strongly associated with HS.[3] We restrict ourselves to nouns, named entities, and hashtags, as we do not observe strong associations for other POS tags. We compute the collocation coefficient *Dice* (Smadja et al. 1996) for each word and inspect the end of the spectrum associated with the HS class.

We observe the – by far – strongest association for the token *#rapefugee*. Furthermore, we perceive strong association for cognates of *rape* such as *rapist* and *rapes*.

To further inspect the influence of these indicators, we compute the probability of their occurrence predicting whether a Tweet is HS or not. We find a probability of 65.8% for *#rapefugee* and of even 87.5% for the group of nouns related to *rape*. When inspecting the Tweets containing those explicit HS indicators, we observe that they are often considered as HS regardless of whether the rest of the Tweet is protective of refugees. Because of this simple heuristic, we remove those Tweets from our data set.

[3] Tokenization is done with Twokenizer (Gimpel et al. 2011) and POS-tagging with Stanford POS-tagger (Toutanova et al. 2003).

3.2 Paraphrasing

To make the Tweets explicit, we paraphrase them according to a set of rules[4], which correspond to previously mentioned paraphrase classes. We apply as many rules as possible to one Tweet in order to make it as explicit as possible. As the corpus is concerned with the refugee crisis, we define *Islam*, *Muslim*, and *refugee* as the targets of HS. If a phrase does not explicitly contain them, we paraphrase it by adding this information as a new subject, object, or adjective or by co-reference resolution. An example for this rule is shown in the first explicit paraphrase:

Im. #Vendetta, #ForcedConversion, #Sharia, #ChildBrides, #Polygamy, #GenitalMutilation - don't see how it belongs to us.
Ex.1 [...] - don't see how **Islam** belongs to us.
Ex.2 [...] - It **doesn't** belongs to us.
Ex.3 [...] - **Islam doesn't** belongs to us.

If the message of the phrase is softened through hedges such as modals (e. g., *could*, *should*) and epistemic modality with first person singular (e. g., *I think*, *in my opinion*) these are either removed or reformulated to be more explicit. This reformulation is shown in the second explicit paraphrase in the example above. However, as we apply as many rules as possible, the Tweet would be paraphrased to its final version as shown in the third paraphrase in the example above. Rhetorical questions are paraphrased to affirmative phrases, e. g.,

- Yesterday the refugees came. Today there's burglary. Coincidence?
- Yesterday the refugees came. [...] **Not a coincidence!**

Furthermore, implicit generalizations are made explicit through the use of quantifiers.

- 90% of all refugees want to come to Germany, only because nobody else will give them money! Islamize in passing. #Lanz
- **All** refugees want to come to Germany, [...].

The paraphrasing process was performed independently by two experts, who chose the same instances of implicit stance, but produced slightly differing paraphrases.

The experts merged the two sets by choosing one of the two paraphrased versions after a discussion.

3.3 Supervised Machine Learning

To examine the influence of implicitness on automatic HS detection, we re-implement a state-of-the-art system. We adapt the systems of Waseem and Hovy (2016) and Warner and Hirschberg (2012) to German data. Thus, we rely on an

[4] https://github.com/MeDarina/HateSpeechImplicit.

SVM equipped with type-token-ratio, emoticon ratio, character, token, and POS uni-, bi-, and trigams features.

For our classification, we consider Tweets as HS in which at least one annotator flagged it as such since we aim at training a high-recall classifier. The resulting class distribution is 33% HS and 67% NO HS. First, we establish baselines by calculating a majority class baseline and conducting a ten-fold cross-validation. We report macro-F_1 for all conducted experiments. While the majority class baseline results in a macro-F_1 of .4, we obtain a macro-F_1 of .65 for the cross-validation.

To inspect the influence of implicitness, we conduct a train-/test-split with the selected implicit Tweets as test instances and the remaining Tweets as train instances. We achieve a macro-F_1 of only .1, regardless whether we use the explicit or implicit version of the Tweets. Although the performance is higher than the majority class baseline, the drop is dramatic compared to the cross-validation.

First, these results indicate that implicitness is a major problem in HS detection and thus should be addressed by future research. Second, as results are the same for the more explicit version, the classifier seems to be incapable of recognizing explicit paraphrases of implicit Tweets. Although this was expected since we did not add HS indicating tokens during paraphrasing, it may be highly problematic as implicitness may alter human perception of HS.

4 User Study

After the exclusion of explicit Tweets, a set of 36 implicit Tweets remained, which were paraphrased into an explicit version. To analyze the difference in their perception, we conducted an online survey using a *between-group* design with implicitness as the experimental condition. The randomly assigned participants had to make a binary decision for each Tweet on whether it is HS and rate its offensiveness on a six-point scale, in accordance with Ross et al. (2016). The participants were shown the definition of *HS* of the European ministerial committee[5].

As understanding the content of the Tweets is crucial, we filtered according to native knowledge of German which resulted in 101 participants. They reported a mean age of 27.7 years, 53.4% considering themselves female, 41.6% male and 1% other genders. 39.6% had a university entrance qualification, 58.4% a university degree, and 1% had another education level. More than 90% stated that they identify as Germans which may question the representativeness of our study. Especially, the educational and ethnic background might be factors strongly influencing the perception of HS. 55 remained in the implicit condition and 46 in the explicit condition.

[5] http://www.egmr.org/minkom/ch/rec1997-20.pdf.

5 Results

First, we inspect how often the Tweets are identified as HS. On average, we find that 31.6% of the Tweets are rated as HS in the explicit ($M_{explicit} = 11.3$)[6] and 40.1% in the implicit condition ($M_{implicit} = 14.4$). Interestingly, we observe a high standard deviation ($SD_{explicit} = 11.3$ and $SD_{implicit} = 14.6$) for both conditions. These findings underline how difficult it is for humans to reliably detect HS and thus align with the findings of Ross et al. (2016). A χ^2 test shows that the answer to this question is not significantly differently distributed in the two conditions, ($\chi^2_{(22,N=57)} = 4.53$, $p < .05$). Regarding intensity, encoded from 1–6, we do not find statistically significant differences between the explicit ($M = 3.9$, $SD = .94$) and the implicit ($M = 4.1$, $SD = .98$) condition according to a T-test ($t(97.4) = 1.1$, $p > .05$). To further analyze this difference, we inspect the difference for each instance, which is visualized in Fig. 1. All except one of the significantly differing instances are perceived as more hateful in the implicit version. For all cases, we observe that the implicit version is more global and less directed, which could be due to the fact that the vague and global formulation targets larger groups. Instances 6 and 10 contain rhetorical questions, which may be perceived as hidden or more accusing than the affirmative rather factual version. The one case in which the explicit form is more offensive is the only instance containing a threat of violence, which becomes more directed through making it explicit.

We also compute the change in the binary decisions between HS and NO HS on the level of individual instances using χ^2. Three of the eight significantly less offensive explicit instances on the scale are also significantly less often considered being HS in the binary decision. Similarly, instance 24, which is perceived significantly more offensive is more frequently considered as HS. Thus, we conclude that there is a relationship between the offensiveness and the HS rating

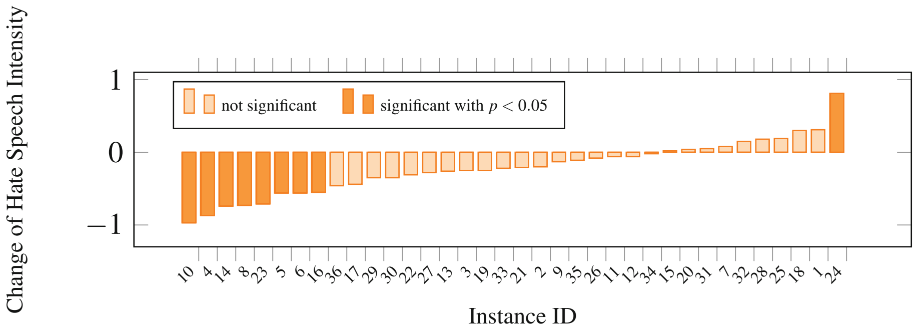

Fig. 1. Change in HS intensity between implicit and explicit versions.

[6] Statistical measures are reported according to the American Psychological Association (1994): M = Mean, SD = standard deviation, p = probability; N = number of participants/annotators.

and that both are equally affected by implicitness. However, the direction of this relationship, is moderated by the contentual factors (e. g., the presence of a threat) which need further investigation.

6 Conclusion and Future Work

In this study we show that there are individual instances of explicit HS which are perceived significantly different compared to their implicit counterparts. However, on average, the polarity of this deviation remains unclear and seems to be moderated by content variables.

In all cases where the implicit version is perceived as more intensely hateful, the Tweets were rather insulting than threatening. The perception change might be due to several reasons: the sly, potentially deceiving nature of implicitness might be perceived as more hateful, whereas the same content expressed clearly might be perceived as more honest and thus less hateful.

Furthermore, although implicitness has an influence on the human perception of HS, the phenomenon is invisible to automatic classifiers. This poses a severe problem for automatic HS detection, as it opens doors for more intense HS hiding behind the phenomenon of implicitness.

Since this study is based on 36 Tweets, the generalizability of the findings may be limited. Thus, in future work a larger study with more data and more fine-grained distinctions between classes such as *insulting* and *threatening content* would give more insight in the correlation between implicitness and HS perception. Additionally it would be interesting to produce implicit paraphrases of explicitly expressed HS and see the effect. Furthermore, more diverse focus groups, such as representatives of diverse religions, origins, and educational backgrounds are required.

Acknowledgements. This work is supported by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group “User-Centred Social Media”.

References

American Psychological Association: Publication Manual of the American Psychological Association. American Psychological Association, Washington (1994)

Bhagat, R., Hovy, E.: What is a paraphrase? Comput. Linguist. **39**(3), 463–472 (2013). ISSN 04194217

Boltužić, F., Šnajder, J.: Back up your stance: recognizing arguments in online discussions. In: Proceedings of the First Workshop on Argumentation Mining, Baltimore, USA, pp. 49–58 (2014)

Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and 2012 International Confernece on Social Computing (SocialCom), Amsterdam, Netherlands, pp. 71–80. IEEE (2012)

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for Twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, Portland, USA, vol. 2, pp. 42–47 (2011)

Hasan, K.S., Ng, V.: Why are you taking this stance? Identifying and classifying reasons in ideological debates. In: Proceedings of the EMNLP, Doha, Qatar, pp. 751–762 (2014)

Madnani, N., Dorr, B.J.: Generating phrasal and sentential paraphrases: a survey of data-driven methods. Comput. Linguist. **36**(3), 341–387 (2010)

Sanchez, L.M., Vogel, C.: IMHO: an exploratory study of hedging in web forums. In: Proceedings of the SIGDIAL 2013 Conference, Metz, France, pp. 309–313. Association for Computational Linguistics, August 2013

Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: detecting stance in tweets. In: Proceedings of the International Workshop on Semantic Evaluation, San Diego, USA (2016, to appear)

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: Semeval-2014 task 4: aspect based sentiment analysis. In: Proceedings of SemEval 2014, pp. 27–35 (2014)

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: aspect based sentiment analysis. In: Proceedings of the 9th SemEval, Denver, Colorado, pp. 486–495. Association for Computational Linguistics (2015)

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: SemEval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th SemEval, San Diego, California, pp. 19–30. Association for Computational Linguistics (2016)

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: Beißwenger, M., Wojatzki, M., Zesch, T. (eds.) Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication. Bochumer Linguistische Arbeitsberichte, Bochum, Germany, vol. 17, pp. 6–9 (2016)

Rus, V., Banjade, R., Lintean, M.C.: On paraphrase identification corpora. In: LREC, pp. 2422–2429 (2014)

Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: a statistical approach. Comput. Linguist. **22**(1), 1–38 (1996)

Sobhani, P., Inkpen, D., Matwin, S.: From argumentation mining to stance classification. In: Proceedings of the NAACL HLT 2015, Denver, USA, pp. 67–77 (2015)

Sood, S.O., Antin, J., Churchill, E.F.: Using crowdsourcing to improve profanity detection. In: AAAI Spring Symposium: Wisdom of the Crowd, vol. 12, p. 06 (2012)

Stefanowitsch, A.: Was ist überhaupt hate-speech. In: Stiftung, A.A. (ed.) Geh sterben. Umgang mit Hate-Speech und Kommentaren im Internet, pp. 11–13 (2014)

Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Sapporo, Japan, vol. 1, pp. 173–180. Association for Computational Linguistics (2003)

Vila, M.M., Martí, A., Rodríguez, H., et al.: Is this a paraphrase? What kind? Paraphrase boundaries and typology. Open J. Mod. Linguist. **4**(01), 205 (2014)

Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of LSM 2012, pp. 19–26. ACL (2012)

Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of NAACL-HLT, pp. 88–93 (2016)

Wojatzki, M., Zesch, T.: Stance-based argument mining - modeling implicit argumentation using stance. In: Proceedings of the KONVENS, Bochum, Germany, pp. 313–322 (2016)

Xu, R., Zhou, Y., Wu, D., Gui, L., Du, J., Xue, Y.: Overview of NLPCC shared task 4: stance detection in Chinese microblogs. In: Lin, C.-Y., Xue, N., Zhao, D., Huang, X., Feng, Y. (eds.) ICCPOL/NLPCC -2016. LNCS, vol. 10102, pp. 907–916. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50496-4_85